Life Cycle Management Considerations of Remotely Sensed Geospatial Data and Documentation for Long Term Preservation

Mohammad Khayat (Adnet Systems Inc) and Steven Kempler (NASA Goddard Space Flight Center)

Abstract:

As geospatial missions age, one of the challenges for the usability of data is the availability of relevant and updated metadata with sufficient documentation that can be used by future generations of users to gain knowledge from the original data. Given that remote sensing data undergo many intermediate processing steps, for example, an understanding of the exact algorithms employed and the quality of that data produced, could be key considerations for these users. As interest in global climate data is increasing, documentation about older data, their origins, and provenance are valuable to first time users attempting to perform historical climate research or comparative analysis of global change. Incomplete or missing documentation could be what stands in the way of a new researcher attempting to use the data. Therefore, preservation of documentation and related metadata is sometimes just as critical as the preservation of the original observational data. The Goddard Earth Sciences – Data and Information Service Center (GES DISC), a NASA Earth science Distributed Active Archive Center (DAAC), that falls under the management structure of the Earth Science Data and Information System (ESDIS), is actively pursuing the preservation of all necessary artifacts needed by future users.

In this paper we will detail the data custodial planning and the data lifecycle process developed for content preservation, our implementation of a Preservation System to safeguard documents and associated artifacts from legacy (older) missions, as well as detail lessons learned regarding access rights and confidentiality of information issues. We also elaborate on key points that made our preservation effort successful; the primary points being: the drafting of a governing baseline for historical data preservation from satellite missions, and using the historical baseline as a guide to content filtering of what documents to preserve. The Preservation System currently archives documentation content for High Resolution Dynamics Limb Sounder (HIRDLS), Upper Atmosphere Research Satellite (UARS), Total Ozone Mapping Spectrometer (TOMS), and the 1960's era Nimbus mission. Documentation from other missions like the Tropical Rainfall Measuring Mission (TRMM), the Ozone Monitoring Instrument (OMI), and the Atmospheric Infra-Red Sounder (AIRS) are also slated to be added to this repository, as well as the other mission datasets to be preserved at the GES DISC.

Introduction

Keeping geospatial data from becoming obsolete requires more than periodically refreshing storage media. Media refresh is necessary to ensure data remains readable and does not succumb to out of date storage technology; however, as data volumes have increased, and geospatial data typically undergoes a number of processing iterations, keeping good records and documentation about these intermediate processing steps, or information about inputs and assumptions, could be just as important as making sure that the bits and bytes are still readable. In the early days of space-based observations, the geospatial community's focus was primarily on gathering data and getting it to principal investigators for their research. Of less concern was the planning for how the observations might be used by future generations or ensuring that the data is useable in the long-term. For these generations of users coming decades after the data was initially collected, proper and complete documents are necessary to gain an understanding of what the original data contents are, their format, the intermediate processing steps used, the exact algorithms employed for processing, and the quality of that data. Given the importance of historical data in climate research and comparative analysis of global climate change over time, having such documentation for first time users of legacy data could be the only link that prevents the collections of our “big data” turning into a big waste or missed opportunities. Therefore, a key consideration for us as stewards of data is to ensure that a baseline is developed for all heritage data; this baseline needs to establishe what is the minimum necessary set of required information that must be preserved with the data to benefit future users.

Through the Earth Observing System Data and Information System (EOSDIS) program (Esfandiari, et. al. 2006), NASA manages data generated by its Earth Observing System (EOS) missions, at 12 Distributed Active Archive Centers (DAACs). EOSDIS grew out of the need to manage the end-to-end stewardship of EOS data, and broaden the access of this data to the increasing number of disciplinary and inter-disciplinary communities. DAACs are responsible for ensuring that data and associated documentation are available to users who employ the data for valuable environmental and climate research both nationally and around the globe. The GES DISC, as one of the EOSDIS DAACs, began archiving data in the early 1990s, with data from the Upper Atmosphere Research Satellite (UARS) and the Total Ozone Mapping Spectrometer (TOMS). GES DISC’s archive increased as new NASA missions generated data. These include atmospheric composition datasets from instruments on the Aura mission: High Resolution Dynamics Limb Sounder (HIRDLS); Microwave Limb Sounder (MLS); and Ozone Monitoring Instrument (OMI), the Aqua mission: Atmospheric Infrared Sounder (AIRS), and more recently the Orbiting Carbon Observatory-2 (OCO-2) instrument. In addition to these, GES DISC also archives precipitation data sets, typified by the Tropical Rainfall Measuring Mission (TRMM) and the recently launched Global Precipitation Measurement (GPM) mission, as well as computational and model data from the Modern Era Retrospective-analysis for Research and Applications (MERRA). These are only listed as a representative set of missions to show the temporal range of data and breadth of disciplines covered [Table 1].

Table 1 –Instruments providing datasets archived at the GES DISC. Given the range of years spanning the early years of space mission, life cycle considerations become critical considerations in preservation of data and ensuing information.

Mission	Launch Year(s)	Comments
Nimbus 1 through Nimbus 7	1964-1978 Multiple Launches	Early metrological research data stored originally on film and magnetic tapes media.
UARS	1991	Upper Atmosphere Research Satellite
TOMS	1996	Total Ozone Mapping Spectrometer
TRMM	1997	Tropical Rainfall Measuring Mission
AIRS (Aqua)	2002	Atmospheric Infra-Red Sounder
OMI (Aura)	2004	Ozone Monitoring Instrument
HIRDLS (Aura)	2004	High Resolution Dynamics Limb Sounder
GPM	2014	Global Precipitation Measurement

OCO-2	2014	Orbiting Carbon Observatory - 2

Preservation artifacts and content is not limited to just data or documents, nor is it made up exclusively of documents created from modern word processing software or the raw data down-linked from the satellite or instruments. Another key point to make is that preservation artifacts also spans a wide spectrum of types ranging from conventional print objects (books and other text objects, geospatial data, images, maps), to modern digital objects (binary data, numeric data sets in ASCII form, video, etc.) Prelaunch calibration data is just as important to preserve as is data from on-orbit phase. Similarly documents describing post processing generation of higher level products are also just as necessary. Content media could also take many forms and, for example, be in the form of analogue film or video, as is the case for example with material from our Nimbus data collection. In the Nimbus case, the GES DISC had to embark on an extensive media refresh to transfer content from data on film and magnetic tapes with 1960's era technologies. As it is becoming harder to find magnetic tape machine readers that are able to read these older media types or simply because of the natural degradation of the storage media, technology refresh is critical to the survival of the original data. For these reasons, proper data management requires that the entire lifecycle of data and information is considered carefully from the creation phase and revisited periodically to ensure that neither the media nor critical metadata becomes unrecoverable.

The NASA Earth Science Division (ESD) now requires all modern missions and projects to develop a Data Management Plan (DMP) to address the management of data from the time data is acquired to their entry into the active archive center for safekeeping as the data is in active use by researchers. Data from NASA funded projects are to be treated as national assets and as such the DMP ensures proper stewardship of the data while it is in active use until plans are made to transfer them (or retire them) to a permanent archive, such as the National Archives, when they are no longer in active use. However, given the level of interest by climate modelers for all data from the earlier satellite based observations, users are still demanding data from early missions like Nimbus and others that have long since ended. DAACs are therefore facing the prospects of supporting these data far longer than originally envisioned. Some of the older missions even predate EOSDIS and the ESD requirements of a Data Management Plan. Basically, preservation of data and related artifacts became a necessity given the overwhelming interest in the data as new users become more adept at using older data creating new products using merged data records for example from multiple instruments. Organizations with data custodial responsibility must therefore be adaptive to respond to the changing landscape of how their data is being utilized. Similarly, as collections grow in complexity and volume the need to establish relationships between data objects in a repository becomes of critical importance to get the full benefit of a structured repository.

One challenge that DAACs face is the continual evolution of their users. Traditional users of NASA DAACs were teams of active scientists, principal investigators, or researchers from university or governmental institutions that were well versed in remote sensing data formats, processing algorithms and tools. Over the last couple of decades, and with the increased accessibility to this data by the public at large using the World Wide Web information portals, DAACs are more and more faced with the need to provide increased tools and services to a multidisciplinary set of users that are not from the traditional earth sciences or remote sensing background. Furthermore in a "big data" era, data scientists are combining and fusing data from across multi-agency data archives to draw new information from historical data or to create new climate models which were not envisioned at the time of the initial mission concept. The GES DISC, for example, routinely serves public users who are accessing remote sensing data for the first time as a matter of curiosity, including students in high school or university only recently exposed to our data. We are also faced with a user base that is increasingly more international and from non-English speaking countries. This magnifies the importance of preserving a full set of data documentation.

Yet another challenge is the preservation of heritage (older) missions before data management plans became a mission requirement. In the absence of a data management plan or a baseline standard for preservation of mission data, a significant amount of work is necessary to identify what needs to be preserved, sorted, tagged, and deposited into database with proper inventory. Without this ground work, we risk either omitting some key artifact from our repository, or overburden it with storing every piece of artifact, even those that are not useful to preserve. This filtering and identifying the minimum set of artifacts that should be preserved is the most time consuming part of the preservation activity for heritage missions not having the benefit of a DMP; a point we will illustrate as we describe our preservation experience further.

To make it possible for future users to draw maximum benefit from the data we archive, it is necessary to not only preserve the original raw and reprocessed data, but to also ensure that all the relevant information including metadata, documentation related to intermediate processing steps or algorithms, important calibration or model input data, instrument design prints, etc. are also preserved as necessary. For example, that there might be "cross-talk" between two bands of an instrument, or a calibration drift over time, are critical information to a user trying to make use of this data years into the future. In this paper we will describe the need for documentation preservation and detail our experiences with preservation implementation for specific missions that the GES DISC has undertaken; we will also detail our implementation for a documentation preservation system and the lessons learned from our experience setting up this repository.

The Case for Document Preservation

NASA's earth observation missions commenced with the Television Infrared Observation Satellite (TIROS) series in the 1960s and continued with the Nimbus and Landsat satellite missions, followed by suites of other observatories. There are now tens of missions that are collecting data about any number of environmental or atmospheric related observations (see: http://science.nasa.gov/earth-science/missions/ for a full listing). NASA's Earth science activities have led to increasingly sophisticated satellite instruments, much larger data volumes, more complex data structures and analyses, and a diverse suite of data products generated with sophisticated data algorithms. NASA now has at its disposal a huge amount of information about the state of our planet obtained from the vantage point of polar and low earth orbit satellites. For scientists seeking to study Earth’s changing climate, having long-term time series of data on key climate variables is crucial. The data from these missions constitute a vital archive for Earth science research.

As many of these NASA EOS missions reach the end of their active life or are nearing it, DAACs are increasingly challenged to ensure that all necessary information that must be conveyed to future users are properly preserved. In order to ensure that future users can draw maximum benefit from this data for years to come, they will require access to documentation about the details of the data, their formats, the processing algorithms or information relating to provenance of the data at various steps. Failing to preserve a document with some key information could hamper a user’s ability to understand anomalies in the data. For example, if information about a calibration drifts of an instrument is not properly captured and conveyed, future users could be at risk of misinterpreting data. This failure could then hamper the data-information knowledge chain [Figure -1] which takes away from the ability of those researchers to draw full value from the data. However, making full use of this information is far from trivial. The past four decades have seen a revolution in information technology as digital data volumes have grown exponentially accompanied with a similar increase in accessibility and interoperability. In the current realm of “big data” the complexity is growing in volume, content, lineage among other factors and creating new professional fields where data scientists with an interdisciplinary approach are looking to use the data for new purposes.

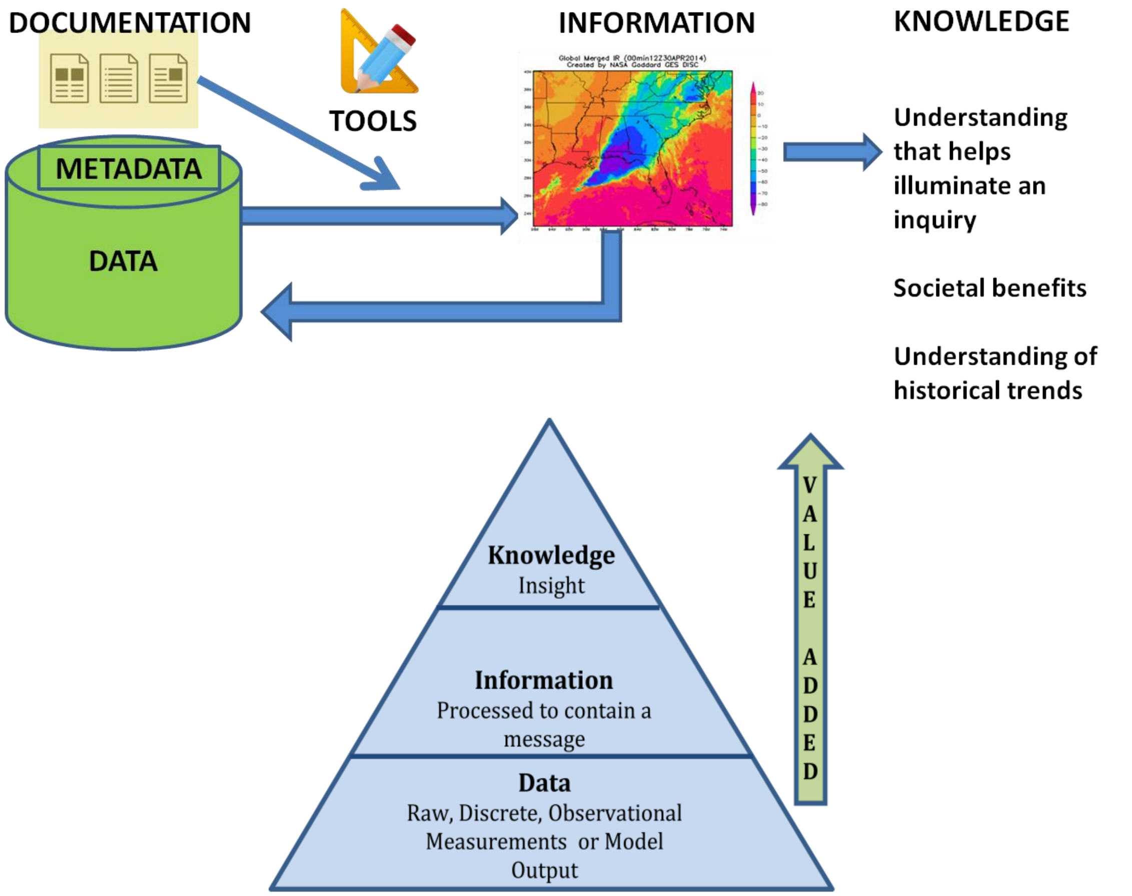

DOCUMENTATION
TOOLS
INFORMATION
KNOWLEDGE
Global Merged IR (00min12Z30APR2014)
Created by NASA Goddard GES DISC
METADATA
DATA
Understanding that helps illuminate an inquiry
Societal benefits
Understanding of historical trends
Knowledge
Insight
Information
Processed to contain a message
Data
Raw, Discrete, Observational Measurements or Model Output
VALUE ADDED

Figure 1. Transition from Data-Information-Knowledge requires traceable inputs and processes that must be preserved so that future users can understand, and glean value from original science data.

The Data-Information-Knowledge Chain is a key reason that drives DAACs to invest in developing the infrastructure to archive, and reprocess data if necessary, to account for periodic refresh of storage media and technology, and in continually identifying what relevant documentation, metadata or other artifacts must be preserved. Each DAAC typically has a well documented process for "*how*" to ensure that the mechanics of data storage and media refresh is performed to ensure the preservation goal. This process addresses the question of how often to refresh storage technology and what media technology is best suited for its mission at any given period. Answering the question of *"what"* must be preserved is another key aspect of archival that requires forethought and the development of a *baseline* for the minimal set that must be preserved and should be a significant aim of a mission DMP. Developing this baseline for older legacy missions was a significant aid in guiding the DAACs to identify the required subset of documentation that needs to be preserved. In the next section, we will elaborate on the baseline guidance developed by EOSDIS which will go a long way to ensuring the relevance of our data well into the future for the benefit of users and as they attempt to derive knowledge from our vast data holdings.

Data Document Preservation at NASA GES DISC

As a DAAC, the GES DISC becomes a designated archive for data generated by selected NASA Earth observing missions. In addition to archiving data, the GES DISC seeks to archive related information, such as algorithm processing details (like Algorithm Theoretical Basis Documents) or interface description documents, science requirement documents, instrument alignment data, instrument calibration data, optical or spectral specifications related to instruments, etc. Unfortunately, on earlier missions prior to 2014, Data Management Plans were not required, so gathering preservation worthy documentation becomes challenging. Fortunately, for missions that the GES DISC is the designated home DAAC for, our resident scientists have developed the experience dealing with these assigned mission given the luxury of time to become familiar with the data, develop necessary processing tools and write description documents that become part of the preserved documentation suite.

For example, the HIRDLS mission that ended in 2008, did not have a Data Management Plan (DMP) to deal with long term preservation of its documentation. As a result, the GES DISC and the National Center of Atmospheric Research (NCAR, the HIRDLS Principle Investigator organization) worked closely together to identify *what* of the many hundreds of artifacts are worthy of long term archive rather than the impractical and costly endeavor to archive everything. This classification and sorting, especially near the end of a mission when human resources are being diverted to other projects, introduces challenges that could easily be eliminated with building a data and information lifecycle process into the mission planning from the outset.

The ESDIS Project recognized the importance of planning for preservation for legacy missions and moved to address the lack of a unified guidance (or baseline) for missions not having a Data Management Plan in place. The “NASA Earth Science Data Preservation Content Specification 423-SPEC-001” (H. K. Ramapriyan, 2013) was developed as a preservation baseline by EOSDIS with input from the various DAACs. The process used the HIRDLS mission as a testing ground for the development of this specification. The current version of this guidance serves as guideline that answers “what” needs to be preserved for a legacy mission. The preservation items categorized in this specification encompass eight different content elements as identified and summarized in [Table 2]. The initial task of the HIRDLS long term preservation involved the sorting of the large number of preservation artifacts and determining what is relevant for preservation. After this determination was made, the next activity involved classifying these items according to one of the eight preservation categories listed in [Table 2]. [Figure-2] provides a pictorial illustration of the process used to identify, sort and tag the many potential items that were candidates for long term preservation. As the figure illustrates, it takes someone knowledgeable about the documents, their content, and relevance to future users to make the determination about justification for long term preservation. After each item is tagged and sorted according to one of the categories of the EOSDIS baseline document, they are stores into an electronic repository that could then be easily made searchable and accessible to external users. One of the reasons for careful sorting and tagging is that some of the documents provided to us for preservation might be subject restrictions because they may contain proprietary material which complicates how we handle these documents in our repository system. The value of the EOSDIS baseline document is in that it provides a necessary guidance for all the various DAACs to complete this preservation activity in somewhat a consistent method and reduce the risk that key artifacts are missed inclusion in the preservation attempt.

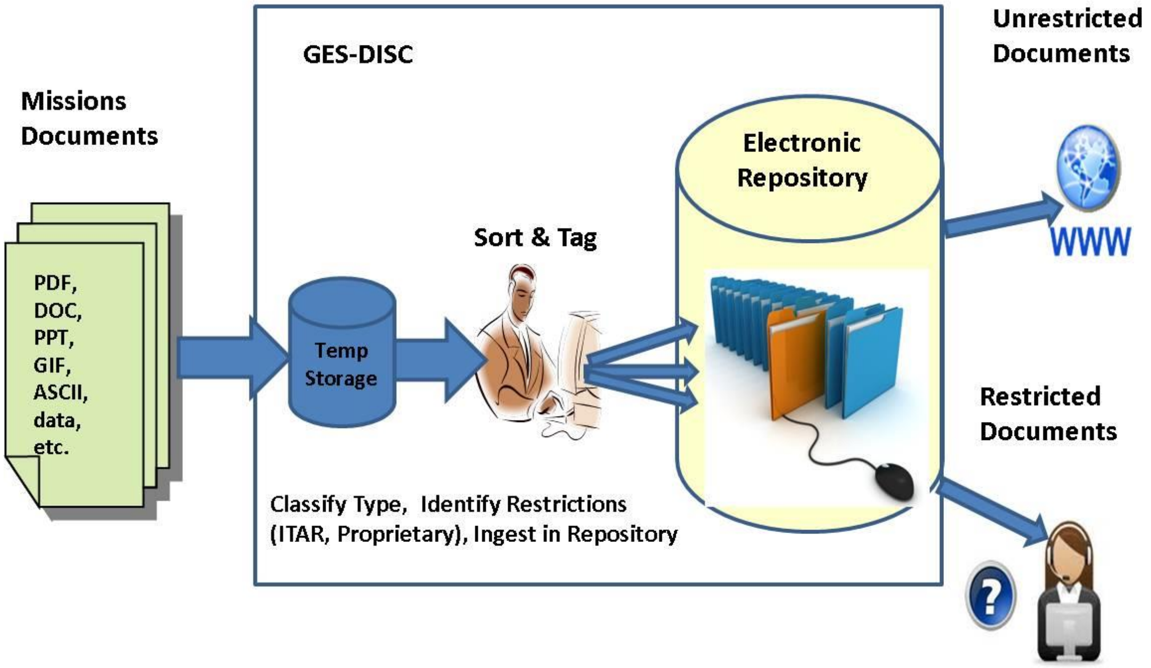

GES-DISC
Missions Documents
PDF, DOC, PPT, GIF, ASCII, data, etc.
Temp Storage
Sort & Tag
Electronic Repository
Unrestricted Documents
WWW
Restricted Documents
Classify Type, Identify Restrictions (ITAR, Proprietary), Ingest in Repository

Figure 2- An overview of the physical objects sorting, tagging, storage process into an archive and distribution system. Restricted documents must be clearly tagged as they require special handling and may not be made accessible publicly.

Table 2- The classification of objects for preservation according to the EOSDIS baseline specification. Artifacts for preservation are classified into one of eight categories as defined by NASA specification document 423-SPEC-001 (H. K. Ramapriyan, 2013).

Preservation Category	Description
Preflight/Pre-Operations Calibration	This element may include instrument specifications, calibration reports, and prelaunch performance measurements.
Science Data Products	This element can include data from the instrument at all processing levels from the Level 0 raw data to Level 3 global and Level 4 model data, as well as metadata required to allow both search and access *for* the data and understanding *of* the data.
Science Data Product Documentation	Many different types of information are included under this data preservation element, including the names of science team members, product requirements, data processing history, algorithm history, detailed algorithm descriptions, and data quality assessment.

Mission Data Calibration	There are two main categories intended for preservation here. One category is descriptions of the calibration methods used for the mission, and the second category is the actual calibration data.
Science Data Product Software	Data collected for this element consists of the software (both description of and the actual code) for the generation of the data product. It is desirable to capture as many different software versions corresponding to the corresponding data product releases as possible
Science Data Product Algorithm Input:	Many remote sensing algorithms require other data (ancillary data) as input to calculate a particular data product. This information includes full descriptions of the input data and attributes covering all input data used by the algorithm, including primary sensor data, ancillary data, forward models (e.g. radiative transfer models, spectral line-lists, optical models, or other model that relates sensor observables to geophysical phenomena) and look-up tables.
Science Data Product Validation	Data types that are classified under this element include the data collected on validation campaigns, accuracy reports, characterization and description of the validation process, ongoing calibration and validation results, and methods used to maintain accurate calibration of the instruments collecting the validation data.
Science Data Software Tools	This often-overlooked (or undervalued) element refers to the tools (mostly software but possibly including hardware) required to read and/or display data collected under the other elements. Data can be in many different formats, requiring specific tools to read and use the data. If these tools are not preserved along with the data, having just the data becomes useless.

Preservation Implementation at the GES DISC

The GES DISC preservation focuses on archiving of binary digital data and supporting documentation from the various NASA funded remote sensing mission or projects. Our concept of operations for user access to documents in our repository is simple access through our main GES DISC system mission portal pages. Users can browse through a catalogue of available documents and download any document that is not subject to restrictions. The first of our instrument datasets to make use of the GES DISC preservation system is HIRDLS.

The physical objects and documents slated for preservation are typically unrestricted documents intended for public distribution. These objects cover all aspects of a satellite remote-sensing mission lifecycle and may include a wide range of content, as described in [Table 2]. Occasionally however, some of the preservation artifacts may contain specific proprietary information (such as manufacturer specific information used in fabrication or design of instruments) or information that is restricted for distribution or subject to the US government import and export restriction (the International Traffic in Arms Regulations -ITAR). Those documents are tagged internally in the preservation system as restricted and are only available to external users by directly contacting the GES DISC User Services. The distribution of these documents is limited and subject to verification of compliance to the applicable access regulations.

For mission data archival, the GES DISC uses an in-house developed system called the Simple, Scalable, Script-Based, Science Product Archive (S4PA) which stores mission data from all levels (Level 0 to Level 4, or higher) (Kempler 2009). In order to facilitate preservation of other artifacts like documentation of various types (.doc, .pdf, .xls, etc.) or hardcopy or film artifacts not suitable for S4PA, we embarked on an effort to setup a document repository system that can catalogue, archive, and distribute documents with low setup costs, but have enough flexibility to meet our evolving data center and user needs. The GES DISC implemented its documentation and digital objects preservation system centered on the open source product Flexible Extensible Digital Object Repository Architecture (Fedora) Commons or Fedora, for short (S. Payette, and C. Lagoze, 1998). Fedora grew out of research that tried to address the needs of university and institutional libraries; these institutions have considerable needs for repository software systems given their expanding digital archiving requirements. That Fedora has the flexibility to address the needs of large repository systems as university libraries gave us some confidence that as an open source system it will have a greater chance of fulfilling our data center needs too.

We chose Fedora as a backbone of our documentation preservation system for its flexible and extensible architecture and available support community built around the open product. Fedora provides the capability to store and archive actual digital copies of items designated for preservation locally on a server at the GES DISC, link to external objects stored at other NASA DAACs, or link to external objects residing within other NASA institutional repositories like the NASA Technical Reports Server (NTRS). Metadata associated with the data products link the associated documentation or other reference items to the mission data products archived in our S4PA data servers. [Figure-3] shows a schematic relationship between the S4PA mission data archive and the Fedora Commons document repository system. Although, the documents are kept in a separate repository database from our S4PA data archive, Fedora easily allows the addition of internal metadata with each artifact object code internally; these additional metadata contain the electronic links that relate each document to the corresponding data by means of dedicated URL links or data object identifiers (DOI) if one exists. The flexibility inherent in the Fedora metadata language allows us to link these documents at the granularity of an entire mission, or specific products by simply adding additional metadata descriptors internally to our repository object identifiers internal to Fedora.

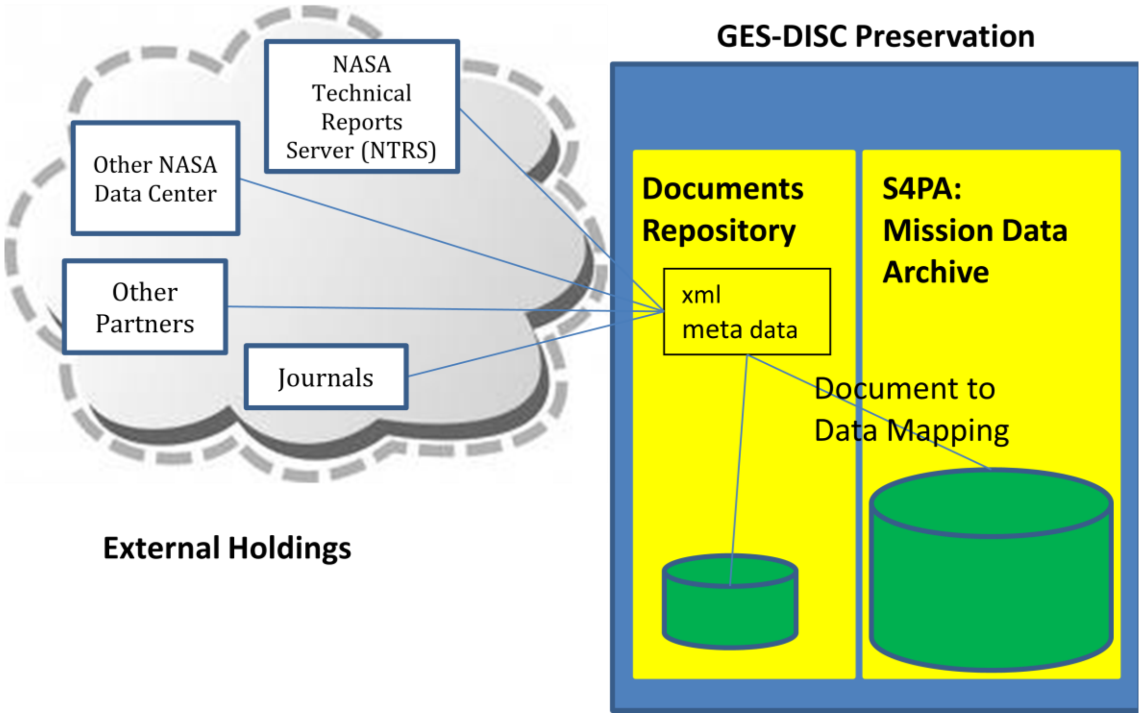


Figure 3- The Fedora Commons based repository works to augment our S4PA archive system for mission data by hosting repository artifacts and providing a mechanism for linking these artifacts to the corresponding mission data.

Although Fedora Commons grew out of the needs of large institutional libraries, we found that it is very well suited for a do-it-yourself (DIY) environment of projects given the wealth of knowledge base available in the community that supports it. With some initial time spent to overcome the learning curve of using this software, we had little trouble to setup a test environment to develop the operational procedures around a test installation. Our rationale for designing the GES DISC preservation documentation repository system based on Fedora Commons software was aided by many of the flexibilities that were inherent in its design. The prime factor in our selection of Fedora Commons is the open source nature of the software, eliminating a costly license purchase or fees for projects with limited funding. it is also supported by a vibrant and active development community, so we can easily find training and reference material for troubleshooting guidance. Furthermore, it is scalable and easy to integrate with other open source applications like Drupal web content management software which makes it easy to integrate with other infrastructure systems at our data center. Also of significant importance to us was that the digital artifacts can be easily ingested into Fedora Commons using a batch ingest method as we typically might receive many documents for bulk ingest into the repository. Finally it is interoperable with semantic capabilities and with a metadata that allows for readily adding descriptors to add relational context amongst the artifacts in the repository, or external content at other data centers.

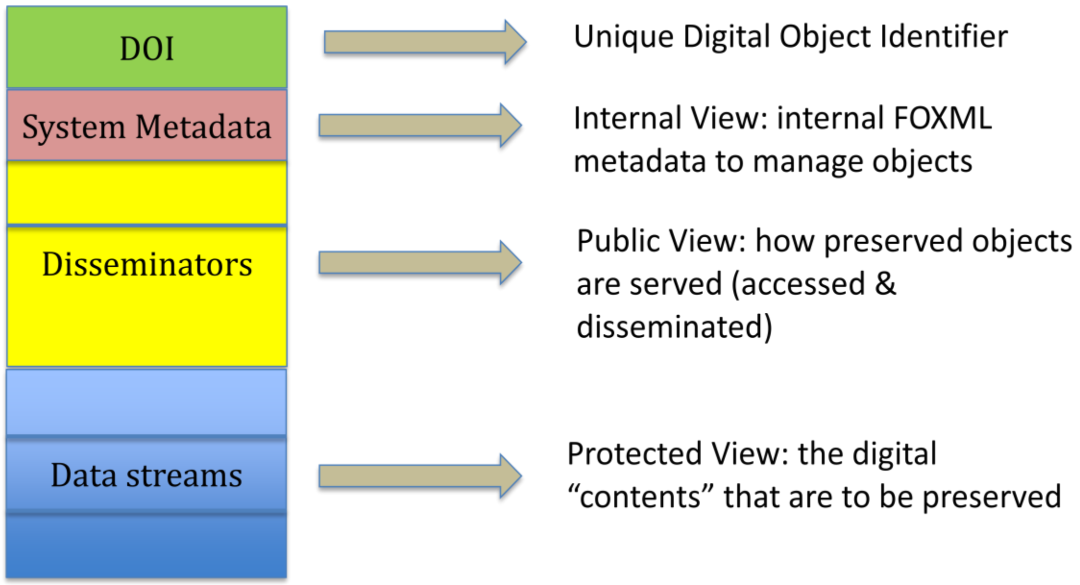
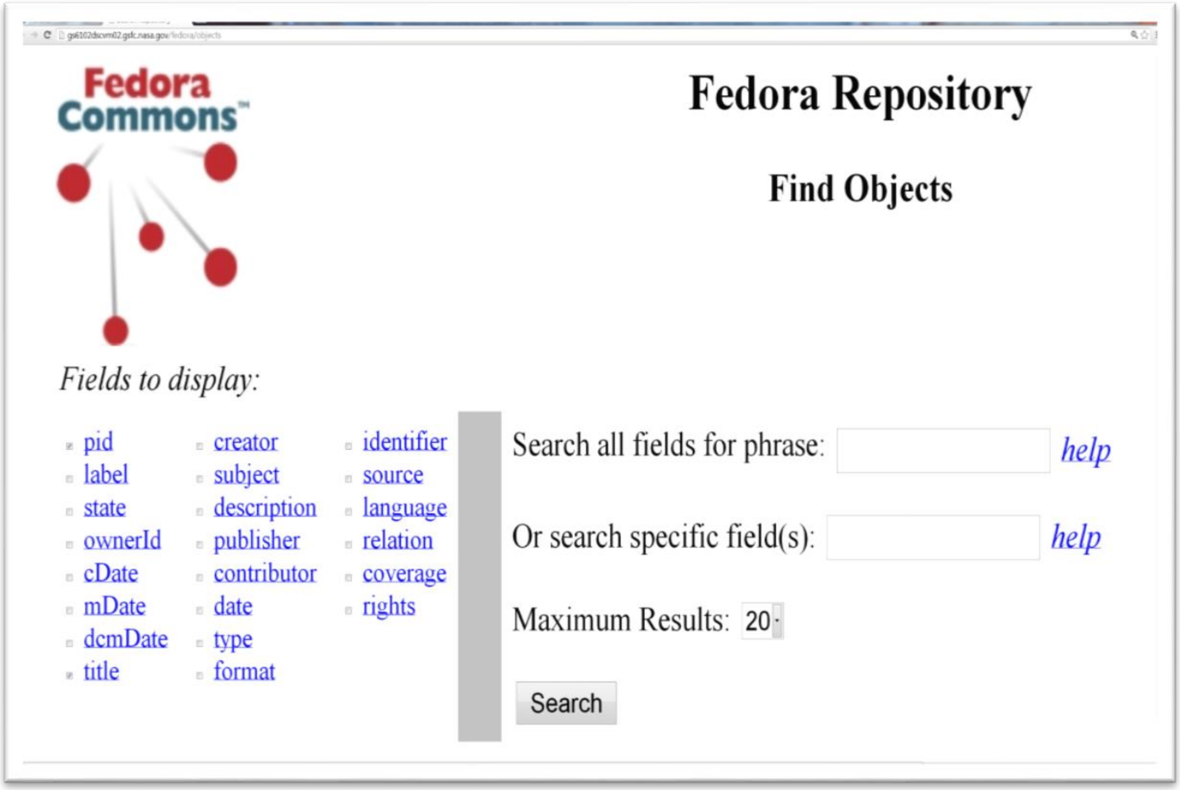

DOI
System Metadata
Disseminators
Data streams
Unique Digital Object Identifier
Internal View: internal FOXML metadata to manage objects
Public View: how preserved objects are served (accessed & disseminated)
Protected View: the digital "contents" that are to be preserved

Figure 4 –Structure of the internal Fedora data structure with object components. The unique identifier is used as reference identification for a preservation item that is stored in the form of an internal Data Stream. The System Metadata and Disseminators are Fedora internal objects to internally store the properties and services for the preserved item.

Fedora provides for an expandable system built on a powerful digital object model which contains an extensible metadata management capability with easy integration of Web services (SOAP and REST). Each object slated for preservation is assigned a persistent identifier or an internal document object identifier (DOI) as a unique digital identifier within the system. Fedora also contains an internal database with an XML (eXtensible Markup Language) like schema (called FOXML) and scripting capability which enables ingest of large quantities of objects more efficiently. Internally Fedora uses intrinsic System Metadata that describes the preservation items and is used only internally to manage the items stored in the repository. Two other components of the internal object structure are the Disseminators and Data Streams. The Data Stream constitutes the item being preserved in the Fedora Repository and is an internal binary structures used to store the digital objects slated for preservation. Disseminators contain the rules, restrictions, and properties for how the preserved item is to be served and presented to the public. It describes internally what regulates the public access and view of the Data Stream. The object structure is further illustrated in [Figure-4] and (S. Payette, and C. Lagoze, 1998).



Fedora Repository

Find Objects

Fields to display:

- ☑ pid
- □ label
- □ state
- □ ownerId
- □ cDate
- □ mDate
- □ dcmDate
- ☑ title
- □ creator
- □ subject
- □ description
- □ publisher
- □ contributor
- □ date
- □ type
- □ format
- □ identifier
- □ source
- □ language
- □ relation
- □ coverage
- □ rights

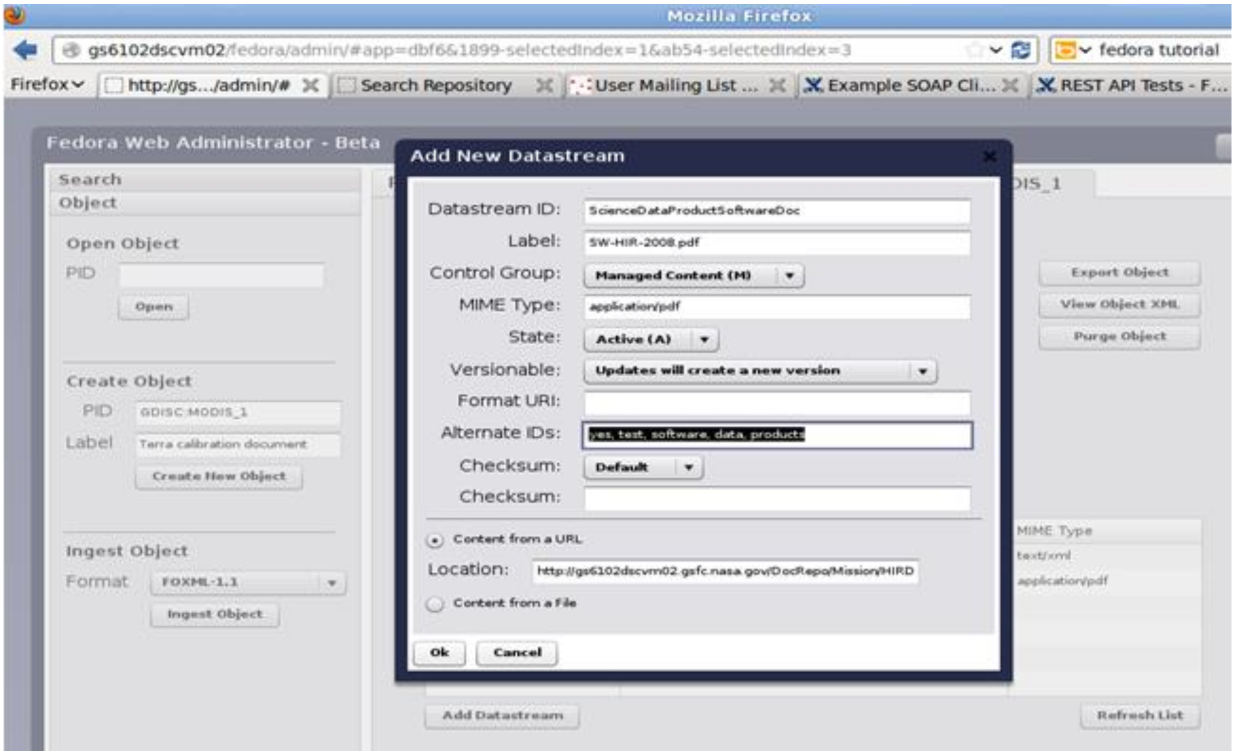
Search all fields for phrase: *help*

Or search specific field(s): *help*

Maximum Results: 20

Figure 5- Fedora Commons provides a simple GUI interface which provide for an easy administration of the system. In addition the interface supports command line and scripted entry which is more convenient for ingesting larger quantity of digital objects with a single operation.

One of the strengths of Fedora is that it has both a command line interface as well as simple and easy to use graphical user interfaces (GUI) which makes both administration of the repository and data entry simple and convenient process, [Figure 5] and [Figure 6]. The command line interface can be used to automate ingest into the repository larger quantities of digital objects within a single batch execution operation. It also provides for version management of the digital objects so that it can keep track of multiple versions of the same document.



PID:

PID: GDISC MODIS_1

Label: Terra calibration document

Format: FOXML-1.1

Datastream ID: ScienceDataProductSoftwareDoc

Label: SW-HIR-2008.pdf

Control Group: Managed Content (M)

MIME Type: application/pdf

State: Active (A)

Versionable: Updates will create a new version

Format URI:

Alternate IDs: yes, test, software, data, products

Checksum: Default

Checksum:

● Content from a URL

Location: http://gs6102dscvm02.gsfc.nasa.gov/DocRepo/Mission/HIRD

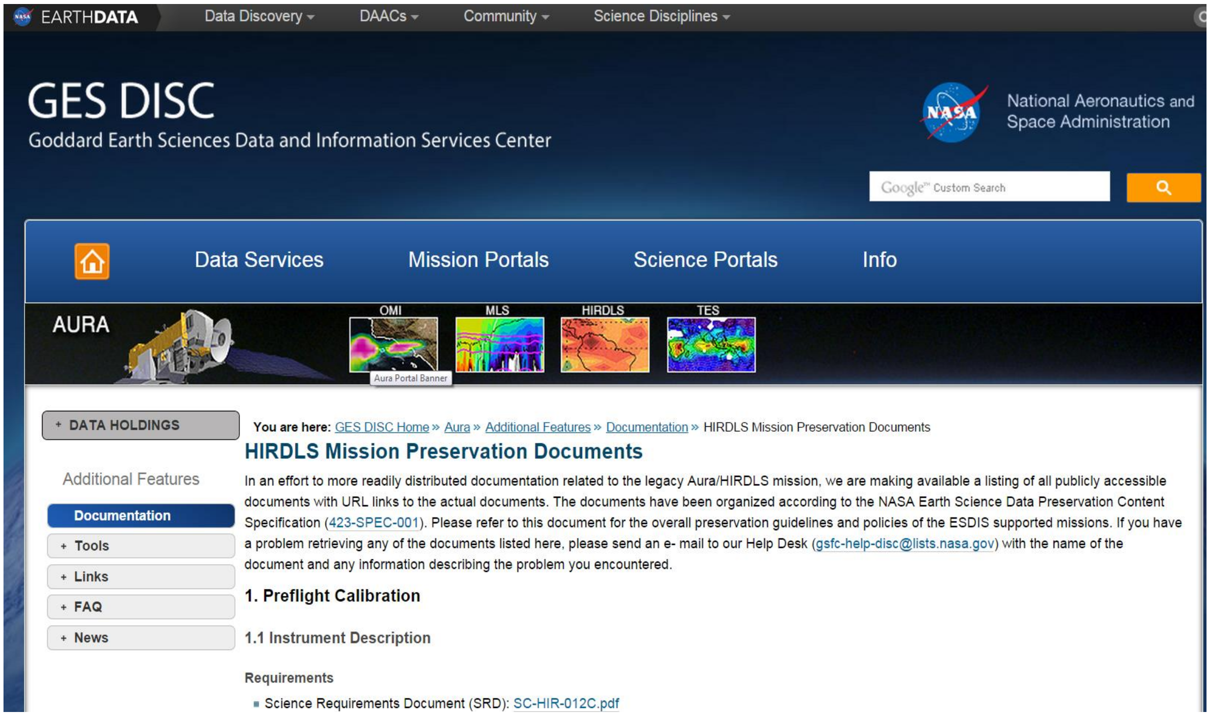
○ Content from a File

MIME Type
text/xml
application/pdf

Figure 6- Fedora Commons provides a simple GUI interface which provide for an easy administration of the system. In addition the interface supports command line and scripted entry which is more convenient for ingesting larger quantity of digital objects with a single operation.

Finally, to tie the repository as a back end engine using Fedora, with a mechanism for users to view the publically available repository content, we set up specific mission portals pages that provide access to these repository documents in an easy to browse format. The documents are presorted and presented according to the classification of the EOSDIS preservation specification. The HIRDLS data preservation portal page (http://disc.sci.gsfc.nasa.gov/Aura/additional/documentation/hirdls-preservation-documents) is exemplified in [Figure-7]. Through this page, users can currently access over four hundred preserved items specific to that mission. Similar access portals provide users with access to over one thousand documents for these heritage missions; the volume of preservation artifacts is growing as the GES DISC receives more items to preserve for these missions.

In summary, to fulfill its obligation to ensure that future users continue to draw benefit from data generated by earlier missions, the GES DISC stood up a digital repository system for all documentation and related artifacts. In this process the GES DISC staff had to self train on the Fedora Commons preservation system using the publicly available documentation by trial and error and through tutorials available online in the public domain. Although there is now an expanding community of users that is making this challenge easier, we found that overcoming local configuration and integration issues can pose challenges that require iteration to resolve, so allocating extra test time for overall project is highly recommended. Creating a dedicated prototype that could be used in a sandbox playground setting to test document ingests and develop web interface connectivity to portals that access Fedora was of great facility to our engineers in their attempts to troubleshoot and eliminate configuration issues.



You are here: GES DISC Home » Aura » Additional Features » Documentation » HIRDLS Mission Preservation Documents

HIRDLS Mission Preservation Documents

In an effort to more readily distributed documentation related to the legacy Aura/HIRDLS mission, we are making available a listing of all publicly accessible documents with URL links to the actual documents. The documents have been organized according to the NASA Earth Science Data Preservation Content Specification (423-SPEC-001). Please refer to this document for the overall preservation guidelines and policies of the ESDIS supported missions. If you have a problem retrieving any of the documents listed here, please send an e- mail to our Help Desk (gsfc-help-disc@lists.nasa.gov) with the name of the document and any information describing the problem you encountered.

1. Preflight Calibration

1.1 Instrument Description

Requirements

- Science Requirements Document (SRD): SC-HIR-012C.pdf

Figure 7- The portal page to access and download the publicly accessible documentation for HIRDLS. Users can view the catalog of available documents from the repository and click to view or download unrestricted documents.

Conclusions and future plans

Because data from NASA's missions are valuable scientific resources of interest to many communities of users, data preservation efforts are intended to allow scientists to continue to use the data well into the future. With every new generation of users comes the possibility of new discoveries and reanalysis of data that was not envisioned earlier. These new and innovative data exploitation methods prolongs the relevancy of heritage missions longer that originally envisioned; it is critical, therefore, to ensure that we understand what information is relevant and take steps to safe keep them.

In the course of performing this preservation task, we identified a number of key factors that are important to ensure that this task is performed properly, namely: creating a baseline for historical data to be preserved, recognizing that preservation artifacts come in many types and formats, making sure that a DMP exists, understanding the relationship between the document or artifact types, and finally that large amount of work is required in the judicious filtering, sorting, and tagging of documents. As documentation constitutes an important mechanism to convey information to new users, deciding what of the many hundreds of formal and informal documents and communications generated in the life of a program is a key consideration and filtering what makes the cut for preservation requires subject matter expertise. In terms of life cycle lessons learned for data and information extraction ability in our preservation activity, we can highlight the following as the most important ones:

- Having a guidelines or baseline governing document helps sift through the many hundreds of potential artifacts to identify what is really worthy of archival is critical. Domain experts then use guidelines to classify each item accordingly, deem it of no value for preservation, or extract what is worthy for preservation into a new document among other alternate approaches.
- There is a wide range of types and formats to documents, just as there is a breadth in storage media types; this is an important aspect of data custodial preservation.
- Open Source software systems could be a cost effective way for establishing an archive system, especially one like Fedora Commons with so many built in functionality suited for preservation; however, one must plan accordingly allowing for sufficient time in project plans to allow for a do-it-yourself project with low preservation budgets. Not having large license expenditures at the outset could quickly be offset with how much longer a DIY project could take to accomplish the task. Although we found a great large amount of online information available in the public domain, Fedora's wealth of functionality takes a good bit of trial and error to setup and configure.
- Lastly, understanding the legal and export limitations surrounding dissemination of information is a complicating factor in preservation of documents. Our DAAC, as is the case for many federal civilian agencies, provides free and unrestricted access to the data we archive. However, this cannot be said about documents as many of them are subject to proprietary commercial restrictions, or subject to U.S. Department of Commerce export control regulations (ITAR). To adequately comply with this regulation for an online repository requires technical considerations (such as user registration and authentication) as well as additional time and effort to understand how to comply with these regulations.

The HIRDLS prototype provided us with the initial experience to initiate a local repository for documents and became the basis for completing the preservation activity for the other missions at the GES DISC. We have extended these services for mission documentations for Nimbus, the Upper Atmosphere Research Satellite (UARS), and Total Ozone Mapping Spectrometer (TOMS). As we continue to expand our repository, we plan to add documents for the Microwave Limb Sounder (MLS), Ozone Monitoring Instrument (OMI), Atmospheric Infrared Sounder (AIRS), among others in the near future. Currently, users can access over four hundred HIRDLS documents online from our portal page. To date, over one thousand documents have been added to our repository for the aforementioned missions and we are continually adding to this digital holding.

One of the challenges still remaining is to setup a user registration that can accommodate the access to preserved documents. Devising an automated user registration mechanism can take proper account of legal considerations and precautions surrounding the handling and distributing these restricted documents. We will continue to monitor our user request metrics to get a better handle on the level of interest for these types of documents and what services to put in place to better serve our users.

References:

1. Esfandiari, M., H. Ramapriyan, J. Behnke, and J. , Sofinowski, 2006, E. "Evolution of the Earth Observing System (EOS) Data and Information System (EOSDIS)", Geoscience and Remote Sensing Symposium. IGARSS 2006. IEEE International Conference on, pages 309 – 312, 2006,(DOI: 10.1109/IGARSS.2006.84) and https://earthdata.nasa.gov/about-eosdis
2. H. K. Ramapriyan, 2013, "NASA Earth Science Data Preservation Content Specification" NASA Goddard Space Flight Center, EOSDIS Project Office Specification (423-SPEC-001), NASA GSFC ***https://earthdata.nasa.gov/sites/default/files/field/document/423-SPEC-001_NASA%20ESD_Preservation_Spec_OriginalCh01_0.pdf***
3. Steve Kempler, Chris Lynnes, Bruce Vollmer, Gary Alcott, and Stephen Berrick, 2009, "Evolution of Information Management at the GSFC Earth Sciences (GES) Data and Information Services Center (DISC)", IEEE Transactions on Geoscience and Remote Sensing, Volume 47, Issue 1, pages 21-28
4. Sandra Payette, and Carl Lagoze, 1998, "Flexible and Extensible Digital Object and Repository Architecture (FEDORA)", European Conference on Research and Advanced Technology for Digital Libraries, Heraklion, Crete, published in Lecture Notes in Computer Science, Springer, 1998, pages 41-59

Acknowledgements

The authors wish to acknowledge extensive support for this work by the following GES DISC staff members supporting this work: Barbara Deshong, Irina Gerasimov, James Johnson, Asghar Esfandiari, and Michael Berganski. In addition, the GES DISC, funded by the Earth Science Data and Information System (ESDIS) Project, is extremely appreciative of their support and guidance.